Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models

Year: 2023 | Citations: 327 | Authors: Emilio Ferrara

Abstract

As generative language models, exemplified by ChatGPT, continue to advance in their capabilities, the spotlight on biases inherent in these models intensifies. This paper delves into the distinctive challenges and risks associated with biases specifically in large-scale language models. We explore the origins of biases, stemming from factors such as training data, model specifications, algorithmic constraints, product design, and policy decisions. Our examination extends to the ethical implications arising from the unintended consequences of biased model outputs. In addition, we analyze the intricacies of mitigating biases, acknowledging the inevitable persistence of some biases, and consider the consequences of deploying these models across diverse applications, including virtual assistants, content generation, and chatbots. Finally, we provide an overview of current approaches for identifying, quantifying, and mitigating biases in language models, underscoring the need for a collaborative, multidisciplinary effort to craft AI systems that embody equity, transparency, and responsibility. This article aims to catalyze a thoughtful discourse within the AI community, prompting researchers and developers to consider the unique role of biases in the domain of generative language models and the ongoing quest for ethical AI.